




1 sourmash: a tool to quickly search, compare, and
2 analyze genomic and metagenomic data sets

3 Luiz Irber ^{1*}, N. Tessa Pierce-Ward ^{1*}, Mohamed Abuelanin ¹, Harriet
4 Alexander ², Abhishek Anant ⁹, Keya Barve ¹, Colton Baumler ¹, Olga
5 Botvinnik ³, Phillip Brooks ¹, Daniel Dsouza ⁹, Laurent Gautier ⁹,
6 Mahmudur Rahman Hera ⁴, Hannah Eve Houts ¹, Lisa K. Johnson ¹,
7 Fabian Klötzl ⁵, David Koslicki ⁴, Marisa Lim ¹, Ricky Lim ⁹, Ivan
8 Ogasawara ⁹, Taylor Reiter ¹, Camille Scott ¹, Andreas Sjödin ⁶,
9 Daniel Standage ⁷, S. Joshua Swamidass ⁸, Connor Tiffany ⁹, Pranathi
10 Vemuri ³, Erik Young ¹, and C. Titus Brown ¹¶

11 1 University of California, Davis 2 Woods Hole Oceanographic Institute 3 Chan-Zuckerberg Biohub 4
12 Pennsylvania State University 5 MPI for Evolutionary Biology 6 Swedish Defence Research Agency (FOI)
13 7 National Bioforensic Analysis Center 8 Washington University in St Louis 9 No affiliation ¶
14 Corresponding author * These authors contributed equally.

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Open Journals](#) 

Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

License

Authors of papers retain copyright
and release the work under a
Creative Commons Attribution 4.0
International License ([CC BY 4.0](#))

15 **Summary**

16 sourmash is a command line tool and Python library for sketching collections of DNA, RNA,
17 and amino acid k-mers for biological sequence search, comparison, and analysis ([Pierce et al.,
18 2019](#)). sourmash's FracMinHash sketching supports fast and accurate sequence comparisons
19 between datasets of different sizes ([Irber, Brooks, et al., 2022](#)), including petabase-scale
20 database search ([Irber, Pierce-Ward, et al., 2022](#)). From release 4.x, sourmash is built on top
21 of Rust and provides an experimental Rust interface.

22 FracMinHash sketching is a lossy compression approach that represents data sets using a
23 "fractional" sketch containing $1/S$ of the original k-mers. Like other sequence sketching
24 techniques (e.g. MinHash, ([Ondov et al., 2015](#))), FracMinHash provides a lightweight way to
25 store representations of large DNA or RNA sequence collections for comparison and search.
26 Sketches can be used to identify samples, find similar samples, identify data sets with shared
27 sequences, and build phylogenetic trees. FracMinHash sketching supports estimation of overlap,
28 bidirectional containment, and Jaccard similarity between data sets and is accurate even for
29 data sets of very different sizes.

30 Since sourmash v1 was released in 2016 ([Brown & Irber, 2016](#)), sourmash has expanded to
31 support new database types and many more command line functions. In particular, sourmash
32 now has robust support for both Jaccard similarity and containment calculations, which enables
33 analysis and comparison of data sets of different sizes, including large metagenomic samples.
34 As of v4.4, sourmash can convert these to estimated Average Nucleotide Identity (ANI) values,
35 which can provide improved biological context to sketch comparisons ([Hera et al., 2022](#)).

36 **Statement of Need**

37 Large collections of genomes, transcriptomes, and raw sequencing data sets are readily
38 available in biology, and the field needs lightweight computational methods for searching and
39 summarizing the content of both public and private collections. sourmash provides a flexible

40 set of programmatic functionality for this purpose, together with a robust and well-tested
41 command-line interface. It has been used in well over 200 publications (based on citations of
42 Brown & Irber (2016) and Pierce et al. (2019)) and it continues to expand in functionality.

43 Acknowledgements

44 This work is funded in part by the Gordon and Betty Moore Foundation's Data-Driven Discovery
45 Initiative [GBMF4551 to CTB].

46 References

- 47 Brown, C. T., & Irber, L. (2016). Sourmash: A library for MinHash sketching of DNA. *Journal*
48 *of Open Source Software*, 1(5), 27. <https://doi.org/10.21105/joss.00027>
- 49 Hera, M. R., Pierce-Ward, N. T., & Koslicki, D. (2022). Debiasing FracMinHash and deriving
50 confidence intervals for mutation rates across a wide range of evolutionary distances.
51 *bioRxiv*.
- 52 Irber, L. C., Brooks, P. T., Reiter, T. E., Pierce-Ward, N. T., Hera, M. R., Koslicki, D., & Brown,
53 C. T. (2022). Lightweight compositional analysis of metagenomes with FracMinHash and
54 minimum metagenome covers. *bioRxiv*.
- 55 Irber, L. C., Pierce-Ward, N. T., & Brown, C. T. (2022). Sourmash branchwater enables
56 lightweight petabyte-scale sequence search. *bioRxiv*.
- 57 Ondov, B. D., Treangen, T. J., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A.
58 M. (2015). Fast genome and metagenome distance estimation using MinHash. *bioRxiv*,
59 029827. <https://doi.org/10.1101/029827>
- 60 Pierce, N. T., Irber, L., Reiter, T., Brooks, P., & Brown, C. T. (2019). Large-scale se-
61 quence comparisons with sourmash. *F1000Research*, 8, 1006. [https://doi.org/10.12688/](https://doi.org/10.12688/f1000research.19675.1)
62 [f1000research.19675.1](https://doi.org/10.12688/f1000research.19675.1)